Research topic: Influence of YouTube video characteristics and creator demographics on viewer engagement and sales performance

General description:

I selected this dataset from Kaggle because it offers comprehensive information on YouTube videos and their creators, consisting of 21 variables for 1096 videos. Collected in 2022 by business analyst interns from KultureHire, the dataset aims to analyze the success ratio of YouTube content creators. Each row corresponds to a different video, with columns detailing various aspects such as video characteristics (e.g., duration, language, subtitles), creator demographics (e.g., total subscribers), and viewer engagement metrics (e.g., number of views, likes, comments). The dataset's thoroughness and lack of known limitations make it ideal for examining factors that contribute to video success on YouTube.

Caution:

I removed potential outliers from the dataset using the 2RMSE method, which excludes observations more than 2 mean square errors from the regression line. After this, 1,069 observations remain.

My main research question:

Which YouTube video and creator characteristics are the strongest predictors of high viewer engagement, suggesting potential sales success?

Understanding the characteristics of a YouTube videos and the creators of those videos that predict high viewer engagement is crucial for marketers and content creators who aim to maximize the impact of their digital marketing strategies. Higher engagement often is directly proportional to increased consumer interest and potential sales success. The findings of this research can also help creators tailor their content to better meet audience preferences.

Hypothesis about the main relationships of interest.

Total Channel Subscribers:

Null Hypothesis: There is no relationship between a YouTube channel's total subscribers and the number of views the video in that channel receives.
Alternative Hypothesis: There is a relationship between a YouTube channel's total subscribers and the number of views the video in that channel receives.

Duration of the video:

Null Hypothesis: There is no relationship between the duration of a YouTube video and the number of views the video receives.
Alternative Hypothesis: There is a relationship between the duration of a YouTube video and the number of views the video receives.

Subtitles:

Null Hypothesis: There is no difference in the number of views received between YouTube videos with and without subtitles.
Alternative Hypothesis: There is a difference in the number of views received between YouTube videos with and without subtitles.

Maximum quality of the video:

Null Hypothesis: There is no relationship between the maximum quality of a YouTube video and the number of views the video receives.
Alternative Hypothesis: There is a relationship between the maximum quality of a YouTube video and the number of views the video receives.

Premiered or not:

Null Hypothesis: There is no difference in the number of views received between premiered YouTube videos and non-premiered YouTube videos.
Alternative Hypothesis: There is a difference in the number of views received between premiered YouTube videos and non-premiered YouTube videos.

Interpretation:

- Rejecting the null hypothesis indicates evidence supporting the alternative hypothesis. For instance, if the null hypothesis for total channel subscribers is rejected, it suggests that changes in subscriber numbers affect viewer engagement (number of views), providing insights into factors influencing engagement and potentially boosting sales success.
- Statistical significance means the observed relationship is unlikely due to chance, indicating that changes in predictors are linked to changes in video views.
- Lack of statistical significance (failure to reject the null) suggests the sample data may be too weak to confirm a relationship between predictors and the outcome, indicating other factors might be influencing the outcome.

Descriptive statistics of the important variables:

Viewer engagement metrics:

	Mean	*Median*	*S.D.*	*Min*	*Max*
NoofLikes	*2.782e+05*	*36000*	*7.033e+05*	*0.000*	*6.400e+06*
NoofComments	*27961*	*1400*	*3.234e+05*	*0.000*	*7.380e+06*
TotalChanelViews	*1.9977e+09*	*3.1525e+08*	*7.5049e+09*	*36.000*	*2.0230e+11*

Creator and video demographics:

	Mean	*Median*	*S.D.*	*Min*	*Max*
TotalChannelSubc~	*1.218e+07*	*2.500e+06*	*2.229e+07*	*34.00*	*2.250e+08*
DurationofVideo	*0.06931*	*0.008194*	*0.1721*	*0.000*	*2.160*
MaximumQualityof~	*1252*	*1080*	*431.6*	*240.0*	*2160*
NoofVideostheChannel	*5498.8*	*443.00*	*25610*	*2.0000*	*4.2000e+05*

Proportions of the important categorical variables (total observations – 1069):

Total channel subscribers (>1,000,000; <= 1,000,000):

- YouTube channels that have more than a million subscribers - 64.08%
- YouTube channels that have less than a million - 35.92%

Subtitle:

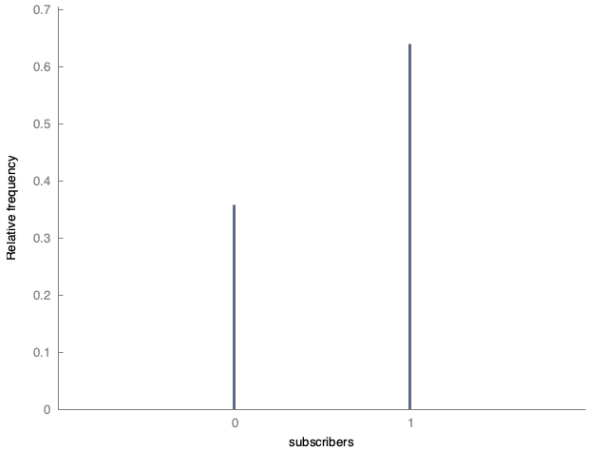
- With subtitles – 60.90%
- Without subtitles – 39.10%

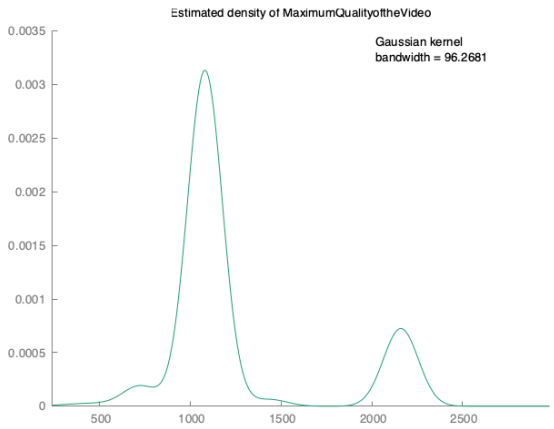
Maximum quality of the video:

- 240px - 0.09%
- 360px - 0.28%
- 480px - 0.47%
- 720px - 4.58%
- 1080px - 75.58%
- 1440px - 1.50%
- 2160px - 17.49%

From the descriptive statistics, we can see that the data consists of around 64.08% of videos posted by creators who have more than a million followers and around 35.92% posted by creators who have less than a million followers. The maximum quality of the video ranges from 240px to 2160px where a majority of those have a maximum quality of 1080px. We are interested in studying any relationships that exist between the viewer engagements and the video characteristics, and the creator demographics variables.

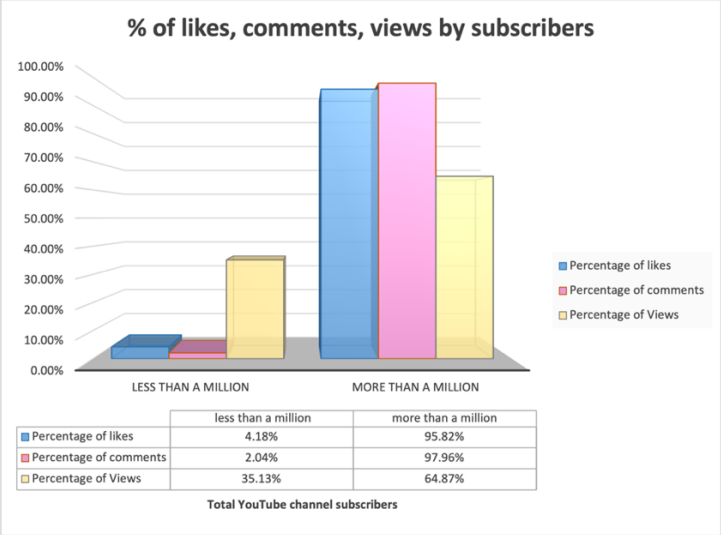
Graphs and density plots:





The frequency distribution of the variable ‘subscribers’ where 0 denotes ‘< 1000000’ subscribers and 1 denotes ‘>1000000’ subscribers. From this sample, we can see that the proportion of videos whose creators have more than a million subscribers are higher than the creators who have less than a million subscribers. The estimated density plot of the variable ‘MaximumQualityoftheVideo’ shows that most number of YouTube videos have a maximum quality of 1080px, as there is a high peak in that area.

% of views, likes and comments by total channel subscribers:



The YouTube videos posted by creators who have more than a million subscribers tends to engage users well compared to the videos posted by creators who have less than a million subscribers.

The relationship between Total Channel Views and Total Channel Subscribers is modeled using the following **quadratic regression equation:**

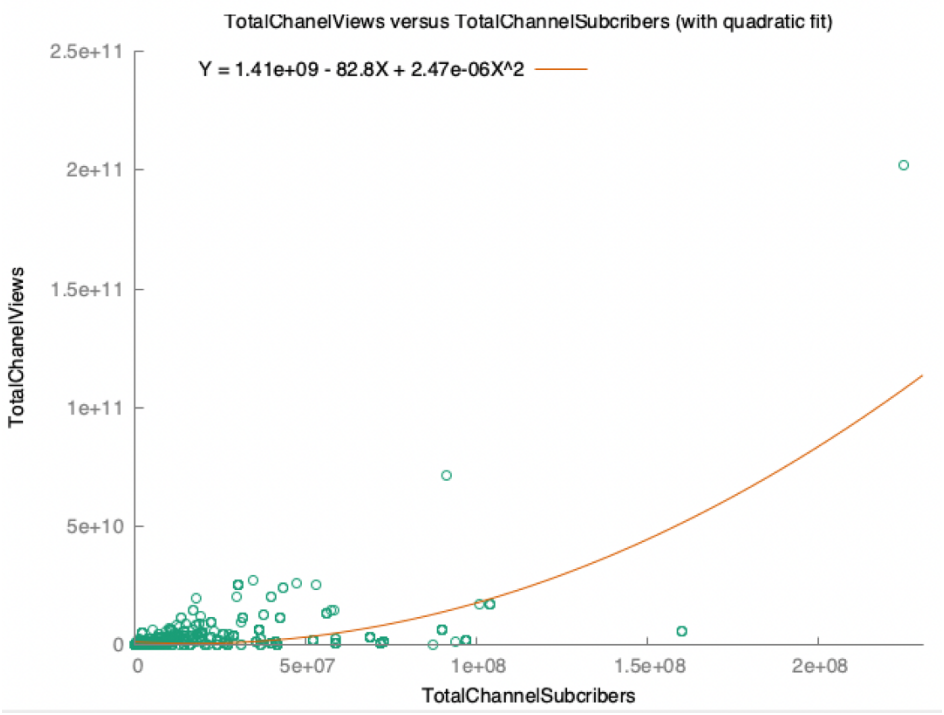
*TotalChanelViews = 1.41e+09 - 82.8(TotalChannelSubcribers) + 2.47e-06(TotalChannelSubcribers*2)*

R-squared = 0.410869

Adjusted R-squared = 0.409764

	coefficient	std. error	t-ratio	p-value	
const	1.41422e+09	2.15077e+08	6.575	7.58e-11	***
TotalChannelSubc~	-82.7529	15.6659	-5.282	1.55e-07	***
Subs2	2.46890e-06	1.36586e-07	18.08	6.52e-64	***

Mean dependent var	2.00e+09	S.D. dependent var	7.50e+09
Sum squared resid	3.54e+22	S.E. of regression	5.77e+09
R-squared	0.410869	Adjusted R-squared	0.409764
F(2, 1066)	371.7223	P-value(F)	3.3e-123
Log-likelihood	-25541.34	Akaike criterion	51088.68
Schwarz criterion	51103.61	Hannan-Quinn	51094.34



Interpretations:

- Initially, gaining more subscribers might slightly reduce the total channel views, as indicated by the -82.8 coefficient. However, as the number of subscribers grows, this negative impact decreases, and eventually, the additional subscribers lead to an increase in total views, thanks to the positive effect of the squared term (2.47e-06).
- This means that the relationship between subscribers and total views isn't straightforward; it starts off negatively but becomes positive as the channel gains more subscribers.

Importance and Policy Implications:

- The findings suggest that while gaining subscribers may initially seem to reduce views, over time, a larger subscriber base leads to more views overall.
- For content creators, this implies the importance of focusing on long-term subscriber growth, as it eventually results in more views.
- Platforms and policymakers should consider supporting tools and strategies that help creators build a larger and engaged subscriber base, as this is key to increasing total channel views over time.

The multiple regression model explores the relationship between Total Channel Views and several predictor variables: Total Channel Subscribers, Duration of Video, Subtitle presence, Maximum

Quality of the Video, Number of Videos the Channel has, and whether the video was Premiered or Not.

TotalChanelViews = -2.49e+09 + 162(TotalChannelSubcribers) + 2.22e+08(DurationofVideo) + 9.16e+08(Subtitle) + 3.41e+05(MaximumQualityoftheVideo + 2.82e+04(NoofVideostheChannel) + 5.88e+08(PremieredorNot)

R-squared = 0.23195

Adjusted R-squared = 0.238919

	coefficient	std. error	t-ratio	p-value	
const	-2.49200e+09	1.12864e+09	-2.208	0.0275	**
TotalChannelSubc~	161.880	9.11619	17.76	5.64e-62	***
DurationofVideo	2.22267e+08	1.19588e+09	0.1859	0.8526	
Subtitle	9.16187e+08	4.15212e+08	2.207	0.0276	**
MaximumQualityof~	341348	467702	0.7298	0.4656	
NoofVideostheCha~	28154.8	7872.74	3.576	0.0004	***
PremieredorNot	5.87531e+08	6.80918e+08	0.8629	0.3884	

Mean dependent var	2.00e+09	S.D. dependent var	7.50e+09
Sum squared resid	4.55e+22	S.E. of regression	6.55e+09
R-squared	0.243195	Adjusted R-squared	0.238919
F(6, 1062)	56.87801	P-value(F)	4.67e-61
Log-likelihood	-25675.21	Akaike criterion	51364.42
Schwarz criterion	51399.24	Hannan-Quinn	51377.61

Statistically Significant Predictors:

- **Total Channel Subscribers**: Positive and highly significant, indicating that more subscribers are associated with higher Total Channel Views.
- **Subtitle (binary)**: Positive and significant, suggesting that videos with subtitles tend to have more views.
- **Number of Videos the Channel has**: Positive and significant, implying that channels with more videos tend to have higher views.

Non-Significant Predictors:

- **Duration of Video**: Not significant, indicating that the length of the video does not have a strong effect on the total views.
- **Maximum Quality of the Video**: Not significant, suggesting that the video quality does not have a notable impact on the total views.
- **Premiered or Not (binary)**: Not significant, implying that whether a video is premiered or not does not significantly affect the total views.

The model highlights the importance of Total Channel Subscribers, Subtitle presence, and the Number of Videos in predicting Total Channel Views, while other factors like Duration of Video, Maximum Quality, and Premiered status do not show a significant impact.

The F-statistic and its corresponding p-value indicate that the overall model is statistically significant. This model can prove to be of value in predicting the factors influencing the YouTube viewer engagement and potential sales success.

Since the variable 'Maximum Quality of the Video' is a categorical variable with multiple categories, a multinomial logit model is an appropriate model to analyze the relationship with the dependent variable 'Total Channel Views'.

	coefficient	std. error	z	p-value	
MaximumQualityoftheVideo = 360					
const	0.974345	1.33968	0.7273	0.4670	
TotalChanelVie~	9.63098e-10	7.26666e-09	0.1325	0.8946	
MaximumQualityoftheVideo = 480					
const	1.05900	1.26530	0.8370	0.4026	
TotalChanelVie~	2.93845e-09	6.66978e-09	0.4406	0.6595	
MaximumQualityoftheVideo = 720					
const	2.99918	1.17014	2.563	0.0104	**
TotalChanelVie~	3.45090e-09	6.63977e-09	0.5197	0.6033	
MaximumQualityoftheVideo = 1080					
const	5.75783	1.16028	4.962	6.96e-07	***
TotalChanelVie~	3.47739e-09	6.63964e-09	0.5237	0.6005	
MaximumQualityoftheVideo = 1440					
const	2.18199	1.19329	1.829	0.0675	*
TotalChanelVie~	3.03643e-09	6.64643e-09	0.4569	0.6478	
MaximumQualityoftheVideo = 2160					
const	4.28166	1.16217	3.684	0.0002	***
TotalChanelVie~	3.48325e-09	6.63964e-09	0.5246	0.5999	

This model indicates that YouTube videos with a maximum quality more than 720px are expected to have higher log-odds of increasing the viewer engagement.

Exploring more nonlinear functional models:

Interaction model:

*TotalChanelViews = 2.45e+07 + 161(TotalChannelSubcribers) + 30.6(TotalChannelSubscribers * DurationoftheVideo)*

R-squared = 0.230401

Adjusted R-squared = 0.228957

```
Dependent variable: TotalChanelViews

                      coefficient    std. error   t-ratio  p-value
  ---------------------------------------------------------------------
  const                2.45104e+07   2.30130e+08   0.1065  0.9152
  TotalChannelSubc~    160.800        9.29448      17.30    2.71e-59 ***
  subs_duration        30.6223       80.9278       0.3784  0.7052

Mean dependent var   2.00e+09   S.D. dependent var   7.50e+09
Sum squared resid    4.63e+22   S.E. of regression   6.59e+09
R-squared            0.230401   Adjusted R-squared   0.228957
F(2, 1066)           159.5684   P-value(F)           2.39e-61
Log-likelihood      -25684.17   Akaike criterion     51374.34
Schwarz criterion    51389.26   Hannan-Quinn         51379.99
```

Log-log model:

Lnviews = 5.08 + 0.974(log(TotalChannelSubscribers)

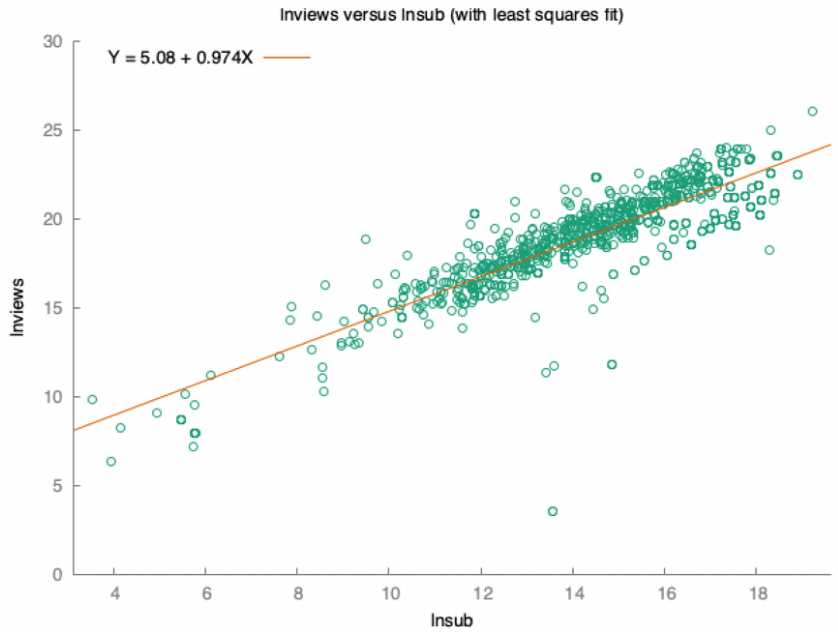
R-squared = 0.742011

Adjusted R-squared = 0.741770

```
Dependent variable: lnviews

            coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const       5.08371      0.258447     19.67    9.66e-74 ***
  lnsub       0.973998     0.0175821    55.40    0.0000   ***

Mean dependent var   19.18627   S.D. dependent var   2.868777
Sum squared resid    2267.595   S.E. of regression   1.457809
R-squared            0.742011   Adjusted R-squared   0.741770
F(1, 1067)           3068.842   P-value(F)           0.000000
Log-likelihood      -1918.787   Akaike criterion     3841.574
Schwarz criterion    3851.523   Hannan-Quinn         3845.343
```



Comparison between the quadratic and the linear log model:

The log-log model explains more of the variation in views (74.2%) compared to the interaction model (23.0%), making it a stronger predictor. The log-log model is simpler, focusing on the relationship between subscribers and views, while the interaction model considers both subscribers and video duration. Overall, the log-log model is better for predicting views, while the interaction model shows how video length might influence the impact of subscribers.

Interpretations:

- The coefficient of 0.974 indicates that for every 1% increase in the total channel subscribers, the number of views increases by approximately 0.974%. This shows a strong positive relationship between the number of subscribers and views.
- The R-squared value of 0.742 means that about 74.2% of the variation in the number of views can be explained by the total number of channel subscribers. The adjusted R-squared is very close at 0.7418, reinforcing that the model is a good fit.

Importance and Policy Implications:

- The high R-squared value suggests that the number of subscribers is a strong predictor of the number of views, meaning that subscriber growth is closely tied to viewership.
- For content creators, this highlights the importance of increasing subscribers as a primary strategy to boost views.
- Platforms and policymakers should support initiatives and tools that help creators attract more subscribers, as this will significantly increase their viewership and overall channel performance.

Important takeaways from this research:

- **Subscriber Growth:** The number of subscribers is a strong predictor of video views. A 1% increase in subscribers leads to approximately a 0.974% increase in views, making subscriber growth essential for boosting viewership and sales potential.
- **Subtitle Presence:** Videos with subtitles tend to have higher views, indicating that adding subtitles can enhance viewer engagement.

- **Number of Videos:** Channels with a larger number of videos generally receive more views, suggesting that consistent content production is important for increasing engagement.
- **Video Quality:** The maximum quality of the video significantly impacts viewer engagement. Higher quality videos, particularly those above 720p, are more likely to increase viewer engagement, making video quality a critical factor for creators to consider.
- **Video Duration and Premiere Status:** These factors did not significantly influence views, indicating that other characteristics, like content quality and relevance, might be more important in driving engagement.
- **Policy Recommendations**: Platforms and policymakers should prioritize tools and strategies that aid creators in increasing their subscriber numbers, as this directly correlates with higher viewership and potential sales success.

Limitations:

One of the creator demographics 'CreatorGender' seems to be inconsistent in this dataset with a lot of values with 'N/A' and blanks. This is a shortcoming so I have not used this in my analysis.

Scope for further analysis:

- **Longitudinal Study on Subscriber Growth:** Expanding the dataset to include temporal data could help analyze how subscriber growth over time influences viewer engagement and sales performance, allowing for a more dynamic understanding of these relationships.
- **Content-Type Segmentation:** Further analysis could segment the data by video content type (e.g., educational, entertainment, product reviews) to identify if certain types of content have stronger correlations with viewer engagement and sales success.

-